A UNIQUE PREDICTION SYSTEM USING CONTENT-BASED FILTERING FOR MOVIES

Parthasarathy G[1], Shanmugapriya D[2]
[1,2] SRM TRP Engineering College, Irungalur, Tiruchirappalli, India.

---***---

Abstract - Internet is an essential part of modern life. Users frequently face the problem of having too much information at their disposal. As users experience the information explosion, recommendation systems (RS) are deployed to help them manage it. In online shopping portals, RS are generally utilised in applications like tourism, entertainment, and e-commerce. In this project, we place emphasis on recreational media, including movies, as a key source of enjoyment and recreation in our daily lives. Because of this, content-based filtering is widely utilised in recommendation systems.

***Key Words*:** Content-based filtering, Movie recommendation, NLP Algorithm

1. INTRODUCTION

Data mining is the process of discovering patterns in large data sets that combines machine learning, statistics, and database systems. It is a subfield of computer science that is interdisciplinary. Data mining is done to uncover and apply relevant information from a dataset. The process of "knowledge discovery in databases" (or KDD) includes data mining as an analysis step [3]. The analysis step of the "Knowledge Discovery in Databases" process, also known as KDD, is data mining, which draws on concepts from computer science and statistics. It uses machine learning, AI, statistics, and systems methods in conjunction [3]. Data mining is intended to transform raw data into a more understandable format that can be further exploited. It encompasses data preparation, model and inference design, interestingness and complexity measurement, structures discovery and post-processing, and visualization [1].

Data mining can be described as a procedure of analyzing large quantities of data to find previously unknown, interesting patterns such as clusters of data records, anomalous records, and dependencies. Typically, databases use spatial indices. This kind of analysis can be used in other ways, such as in predictive analytics or machine learning. In other words, data mining may detect multiple groups within the data, which can then be used by a decision support system to produce more accurate prediction results. A KDD project has three stages, data collection, data preparation, and result interpretation and reporting, but they are steps that aren't specifically called out as part of the KDD process, but rather are included in the overall process.

2. LITERATURE SURVEY

A theoretical model and a short video of the proposed RS-integrated interactive movie were presented in this paper. This suggested method creates a tool for making adaptive movie suggestions for online community systems. Included is a new approach to social networks which is thought to be capable of controlling the dynamics of information exchanged in groups. This proposed replica will perform user methodical analysis and provide a compelling presentation of how social networks shift quickly and frequently[5].

Took a look at some of the problems which can occur with the presentation of recommendations in the movie domain. In this research, we will investigate other previous techniques, which utilize user opinion and approval, and some of the most commonly used RS that utilize these techniques. This paper compares various methods. Use approval and client satisfaction are most effective when they are "planned outline" and "textbook and videotape" linked, as a significant constructive association between customer satisfaction and approval was also engendered in all research studies[6].

Several approaches were examined for RS. CF, Hybrid Recommendations, and Content-Based Recommendation approaches may be classified into three categories. In addition, this paper talks about the benefits and drawbacks of recommendation approaches. Another problem discussed in this paper is that problems can occur in RS too[7].

Created the design of the RS and used four datasets to construct it. The included components include crawling PDFs, generating use models, and others. Additionally, the architectural design, which utilizes the content-based recommendation method for calculating purpose, has been used[8].

The presentation has introduced novel concepts that assist in making information-based decisions. Recognizing which relevant information gives discrepancy ratings also depends on statistical information. To test how well the programme works, we run a simulation of the real dataset containing 12 pieces of contextual information. In simulation, the simulation result revealed that, relative to the assessment method found to be inappropriate, there was a marked difference in the prediction of ratings that was found to be appropriate by this approach and one that was found to be inappropriate, and this showed the significance of the power study and the merits of the presented method in the case of a small dataset[9].

3. CONTENT-BASED FILTERING

Based on attributes, recommendations are made using content-based filtering[4].

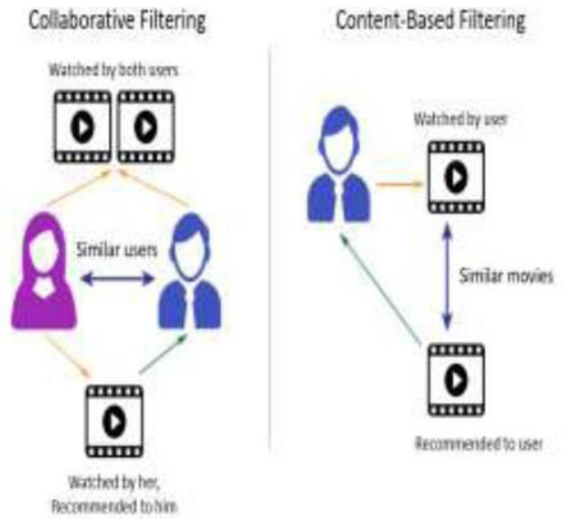

Fig -1: Content Based Filtering

Content-based recommendation engines, which is the movie recommendation system we developed, utilises various attributes or characteristics of products/movies, for example genre, cast, director, and keywords to arrive at content-based recommendations for users. Next, a similarity matrix is calculated by ranking other products and movies based on how similar they are to the product and movie that the individual has liked[4,10].

When new users or new items are introduced, there can be a "cold start" problem because there is insufficient data for collaborative filtering to work more precisely. In order to recommend newly-added items to users who have similar tastes, a large number of users must give the items a rating. The recommendation algorithm doesn't rely on item problem because the algorithm is content-based, not impacted by ratings. And before a new item can be recommended, a large number of users must rate it in a collaborative filtering system. Here, "i.e." is used to describe predictions which are frequently inaccurate if the user is a newcomer, or has not yet watched any movies[4].

The word “Amazing” is present 2 times and the word “Spiderman” is present 1 time[4]. The word “Amazing” is present 1 time and the word “Spiderman” is present 2 times. Now, let’s go and plot this on a 2-Dimensional graph[1]. Text from Movie A will have the point (1,2) and The Text from Movie B will have the point (2,1) where the X-axis on the graph indicates the number of times the word “Spiderman” appears and the Y-axis indicates the number of times word “Amazing” appears[1]. The origin point for both vectors is (0,0). We can change text to a similar vector of word counts by using a Count Vectorizer function or just by doing what we did above[1].

Now, the two texts are represented as vectors and the closer the vectors angular distance are, the more similar they are[10]. So, we can simply get the angular distance which is called theta and represented by the symbol θ to find the similarity between the two vectors[10]. When thinking in terms of probability, machine learning, and likelihood it makes even more sense to use cos θ to get the similarity of the two vectors, this ensures that the value returned is between 0 and 1 since cos 90° = 0 and cos 0° = 1[10].

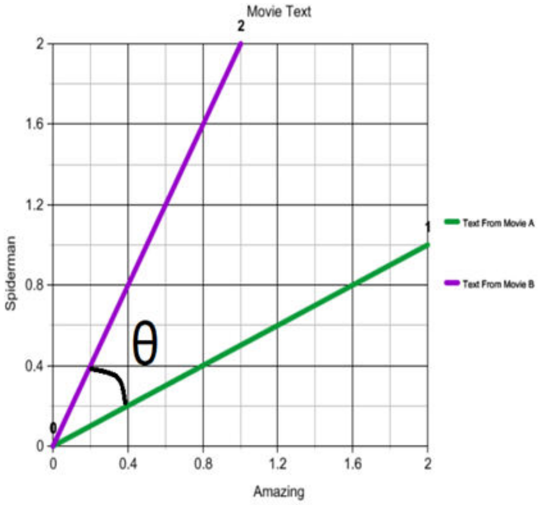

Fig -2: Text graphed as vectors

Now we understand how to get similarities in 2-Dimensions for text represented as vectors and this method can be used for N-Dimensions as well where N is an arbitrary positive integer[1]. So, in summary, we can get the similarity of text by changing the text into vectors and getting the angular distance (θ) between values 0 and 1 using cos θ and ultimately getting a similarity value between 0 and 1[1].

The recommender model can only read and compare a vector (matrix) with another, so we need to convert the ‘Bag_of_words’ into vector representation using Count Vectorizer, which is a simple frequency counter for each word in the ‘Bag_of_words’ column[2]. Once I have the matrix containing the count for all words, I can apply the cosine similarity function to compare similarities between movies[2].

The Top 10 Recommended Movies B the System .The movie recommendation system that we created provides good predictions based on the data set used for training and testing the model[10]. The input given to the recommendation system from the data set and in the output, it provides the most similar movies that the user should watch next[10]. The system can only compare vector so the” bag of words” are converted in to vector using Count Vectorizer. Which is a simple frequency counter for each word in “bag of words” columns which is the used to find the cosine_ similarity Matrix[10]. Here, the similarity between the Movies are found via natural language processing using cosine similarity Matrix[10].

4 METHODOLOGY

4.1 Data Sources

Database, data warehouse, World Wide Web (WWW), text files and other documents are the actual sources of data. You need large volumes of historical data for data mining to be successful[3]. Organizations usually store data in databases or data warehouses. Data warehouses may contain one or more databases, text files, spreadsheets or other kinds of information repositories. Sometimes, data may reside even in plain text files or spreadsheets[3]. World Wide Web or the Internet is another big source of data[3].

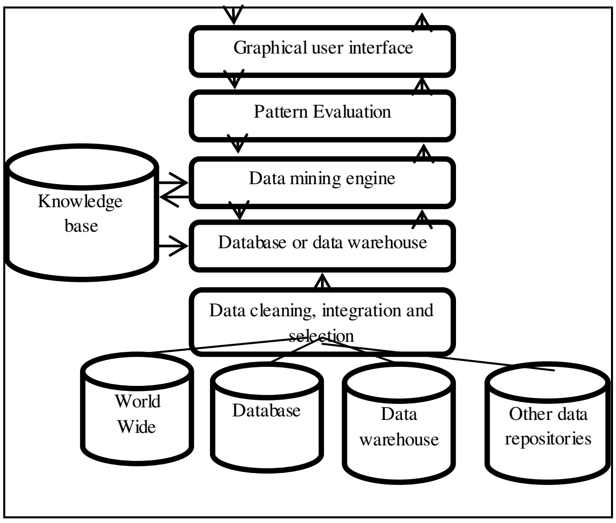

Fig -3: Architecture of Data Mining

The data needs to be cleaned, integrated and selected before passing it to the database or data warehouse server[3]. As the data is from different sources and in different formats, it cannot be used directly for the data mining process because the data might not be complete and reliable[3]. So, first data needs to be cleaned and integrated[3]. Again, more data than required will be collected from different data sources and only the data of interest needs to be selected and passed to the server. These processes are not as simple as we think[3]. A number of techniques may be performed on the data as part of cleaning, integration and selection.

4.2 Database or Data Warehouse Server

The database or data warehouse server contains the actual data that is ready to be processed[3]. Hence, the server is responsible for retrieving the relevant data based on the data mining request of the user[3].

4.3 Data Mining Engine

The data mining engine is the core component of any data mining system[3]. It consists of a number of modules for performing data mining tasks including association, classification, characterization, clustering, prediction, time-series analysis etc[3].

4.4 Pattern Evaluation Modules

The pattern evaluation module is mainly responsible for the measure of interestingness of the pattern by using a threshold value[3]. It interacts with the data mining engine to focus the search towards interesting patterns[3].

4.5 Graphical User Interface

The graphical user interface module communicates between the user and the data mining system[3]. This module helps the user use the system easily and efficiently without knowing the real complexity behind the process[3]. When the user specifies a query or a task, this module interacts with the data mining system and displays the result in an easily understandable manner[3].

4.6 Knowledge Base

The knowledge base is helpful in the whole data mining process. It might be useful for guiding the search or evaluating the interestingness of the result patterns[3]. The knowledge base might even contain user beliefs and data from user experiences that can be useful in the process of data mining. The data mining engine might get inputs from the knowledge base to make the result more accurate and reliable. The pattern evaluation module interacts with the knowledge base on a regular basis to get inputs and also to update it[3].

4.7 DFD LEVEL 0

The Level 0 DFD shows how the system is divided into 'sub-systems' (processes), each of which deals with one or more of the data flows to or from an external agent, and which together provide all of the functionality of the system as a whole[13]. It also identifies internal data stores that must be present in order for the system to do its job, and shows the flow of data between the various parts of the system.

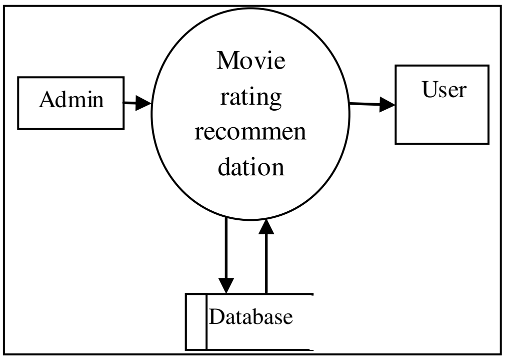

Fig -4 Level 0 Data Flow Diagram

4.8 DFD LEVEL-1

The next stage is to create the Level 1 Data Flow Diagram. This highlights the main functions carried out by the system[7]. As a rule, to describe the system was using between two and seven functions - two being a simple system and seven being a complicated system. This enables us to keep the model manageable on screen or paper.

5 EXPERIMENT AND RESULTS

The recommendation system is part of routine life where people rely on knowledge for deciding their interests[13]. The collaborative filtering model takes data from a user's previous behavior[1] (i.e., previously purchased items or chose or numerical ratings provided to the items) as well as similar decisions made by other users[8].

5.1*Cleaning and Preprocessing*

Data cleansing or *data cleaning* is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the *data* and then replacing, modifying, or deleting the dirty or coarse *data[5]*. In computing, stop words are words which are filtered out before or after processing of natural language data (text).

5.2 *Deep Learning Process*

Deep Learning algorithms have great potential for research into the automated extraction of complex data representations[9]. Deep Learning algorithms can develop a layered, and hierarchical architecture of learning and representing data. Deep Learning Big Data allows extraction of high-level, complex abstractions as data representations through a hierarchical learning process[9]. A key benefit of Deep Learning is Big Data analysis that it can learn from massive amounts of unsupervised data.

5.3 Sentiment Analysis

Using NLP, statistics, or machine learning methods to extract, identify, or otherwise characterize the sentiment content of a text unit[6]. Sometimes referred to as opinion mining, although the emphasis in this case is on extraction[8].Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral[11].

We used the movie dataset publicly made available by twitter API[5]. Analyses were done on this labeled datasets using various feature extraction technique. We used the framework where the preprocessor is applied to the raw sentences which make it more appropriate to understand[6]. Further, the different machine learning techniques trains the dataset with feature vectors and then the semantic analysis offers a large set of synonyms and similarity which provides the polarity of the content[5].

Our proposed work is divided into (i) Upload movie datasets (ii) Read the datasets (iii) Predict user rating and similarity matrix construction (iv) MPL model set construction (v)Recommend the movies and (vi)Performance Evaluation

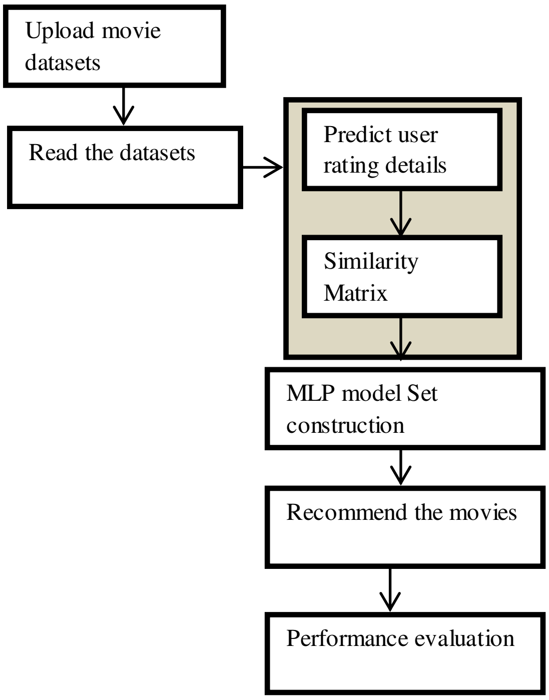

Fig-5 *System Architecture*

We can evaluate the performance using accuracy metrics[2]. The accuracy metric is evaluated as

$$\text{Accuracy}=(TP+TN)/(TP+TN+FP+FN) \qquad (1)$$

The proposed algorithm provide improved accuracy rate than the machine learning algorithms.

Accuracy (ACC) is found as the fraction of total number of perfect predictions to the total number of test data.It can also be represented as 1 – ERR. The finest possible accuracy is 1.0,whereas the very worst is 0.0.

$$ACC=\frac{TP+TN}{TP+TN+FN+FP}*100 \qquad (2)$$

The experiment has conducted using the various models and their performance is shown in Figure 6. Out of the five models multilayer perceptron has better accuracy. It shows 85% of accuracy, which is greater than 10%.

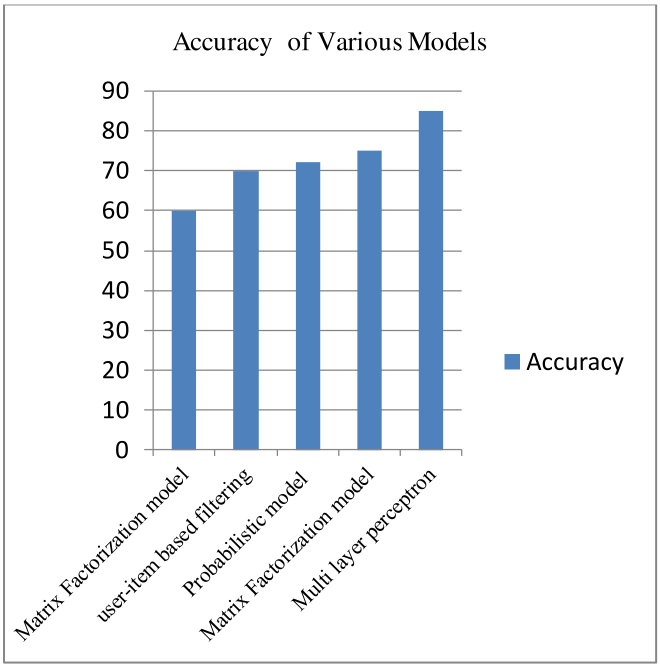

Fig-6 *Performance measures*

6 CONCLUSION

This work presents a novel application of deep learning to a movie recommendation system. Our method provides the following benefits: A visually organized database displays the underlying structure, and an appreciable reduction in the number of search results needed for each result. Using a subset of the movies liked by each user as input to the system, we conducted an in-depth evaluation of our method. In this experiment, the results far outperform a random approach. Nevertheless, we believe we can do better if we have better data and if we make various improvements to our method. An essential component of the system is a deep learning algorithm that deals with the shortcomings of traditional collaborative and content-based recommendations. Multiple approaches are available. Using a hybrid system would allow the inclusion of data specific to the user, like information about the movies their friends like, and content from reviews written by the user and their degree of similarity to other interests and the like.

REFERENCES

1. Mr.Omprakash Yadav,Krishna Mishra, DhananjayPatil, ElvisBraganza.”Design and Implementation of Movie Recommendation System Based On NLP And Content Based Filtering Algorithm” . IRJET, June 06,2020.
2. Wu, Fangzhao, and Yongfeng Huang. "Sentiment domain adaptation with multiple sources." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1. 2016.
3. AhdiRamadhani,Fatma Indriani,Dodon T.Nugrahadi .”Comparison of Navie Bayes Smoothing Methods for Twitter Sentiment Analysis”.12,December 2017.
4. Anand Shanker Tewari, Jyoti Prakash Singh.” Generating Top – N Items Recommendation Set Using Collaborative, Content Based Filtering And Rating Variance. International Conference on Computational Intelligence and Data Science(ICCIDS 2018).
5. Kim, Mucheol, and Sang Oh Park, "Group affinity based social trust model for an intelligent movie recommender system", Multimedia tools and applications 64, no. 2, 505-516, 2013
6. Nanou, Theodora, George Lekakos, and KonstantinosFouskas, "The effects of recommendations‟ presentation on persuasion and satisfaction in a movie recommender system", Multimedia systems 16, no. 4-5, 219-230, 2010
7. Sharma, Meenakshi, and Sandeep Mann, "A survey of recommender systems: approaches and limitations", Int J InnovEng Technol. ICAECE-2013, ISSN , 2319-1058,2013.
8. Beel, Joeran, Stefan Langer, BelaGipp, and AndresNürnberger, "The Architecture and Datasets of Docear's Research Paper Recommender System", D-Lib Magazine 20,no. 11, 2014

9. Ante, Marko Tkalcic, Jurij F. Tasic, and Andrej Kosir, "Predicting and detecting the relevant contextual information in a movie-recommender system", Interacting with Computers, 2013.
10. Emma Grimaldi,”How to build a Content-Based Movie Recommender System With Natural Language Processing”,Oct.1.2018.
11. Mouthami, K., K. Nirmala Devi, and V. MuraliBhaskaran. "Sentiment analysis and classification based on textual reviews." Information communication and embedded systems (ICICES), 2013 international conference on. IEEE, 2013.
12. Cho, Sang-Hyun, and Hang-Bong Kang. "Text sentiment classification for SNS-based marketing using domain sentiment dictionary." Consumer Electronics (ICCE), 2012 IEEE International Conference on. IEEE, 2012.
13. Vinodhini, G., and R. M. Chandrasekaran. "Sentiment analysis and opinion mining: a survey." International Journal 2.6 (2012): 282-292.
14. Khan, Aurangzeb, and BaharumBaharudin. "Sentiment classification using sentence-level semantic orientation of opinion terms from blogs." National Postgraduate Conference (NPC), 2011.IEEE, 2011.
15. Pan, SinnoJialin, et al. "Cross-domain sentiment classification via spectral feature alignment." Proceedings of the 19th international conference on World wide web. ACM, 2010.
16. W.Goa and F.Sebastiani,”Tweet Sentiment: From Classification to Quantification,” in Proc.IEEE/ACM Int.Conf.Adv.Social Netw.Anal.Mining (ASONAM),Aug.2015,pp.97_104.
17. K.H.-Y.Lin,C.Yang, and H.-H.Chen,“What Emotions do News Articles Trigger in their Readers?” in Proc.ACM SIGIR, Jul.2007,pp.733_734.
18. K.H.-Y.Lin,C.Yang, and H.-H.Chen, “Emotion Classification Of Online News Articles From the Reader’s Perspective,” in Proc.IEEE/WIC/ACM WIIAT,Vol.1. Dec.2008,pp.220_226.
19. L.Ye,R.Xu, and J.Xu, “Emotion Prediction Of News Articles From Reader’s Perspective Based On Multi-Label Classification”, in Proc.Int. Conf. Mach.Learn.Cybern.,Vol.5.Jul.2012,pp. 2019_2024.
20. W.B.Liang,H.C.Wang,Y.A.Chu, and C.H.Wu, “Emotion Recommendation in Microblog using Affective Trajectory Model,” in Proc.Asia Pacific Signal Inf. Proc. Assoc.Ann. Summit conf.(APSIPA), Dec.2014,pp.1_5.